Scalable Joint Modeling of Longitudinal and Point Process Data for Disease Trajectory Prediction and Improving Management of Chronic Kidney Disease

Joseph Futoma
Dept. of Statistical Science
Duke University
Durham, NC 27707

Mark Sendak
Institute for Health Innovation
School of Medicine
Duke University
Durham, NC 27707

C. Blake Cameron
Division of Nephrology
Duke University
Durham, NC 27707

Katherine Heller
Dept. of Statistical Science
Duke University
Durham, NC 27707

Abstract

A major goal in personalized medicine is the ability to provide individualized predictions about the future trajectory of a disease. Moreover, for many complex chronic diseases, patients simultaneously have additional comorbid conditions. Accurate determination of the risk of developing serious complications associated with a disease or its comorbidities may be more clinically useful than prediction of future disease trajectory in such cases. We propose a novel probabilistic generative model that can provide individualized predictions of future disease progression while jointly modeling the pattern of related recurrent adverse events. We fit our model using a scalable variational inference algorithm and apply our method to a large dataset of longitudinal electronic patient health records. Our model gives superior performance in terms of both prediction of future disease trajectories and of future serious events when compared to non-joint models. Our predictions are currently being utilized by our local accountable care organization during chart reviews of high risk patients.

1 INTRODUCTION

With the dawn of precision medicine and accountable care, it will become increasingly important for healthcare organizations to make accurate predictions about individual patients' future health risks to improve quality and contain costs. Accountable care organizations (ACOs) are organizations that bear financial responsibility for the quality and total cost of healthcare services provided to a defined population of patients. In order to deliver the right care at the right time in the right setting, ACOs need personalized prediction tools that identify individual patients in their populations at greatest risk of having poor clinical outcomes [Parikh et al., 2016, Bates et al., 2014]. Most ACOs currently lack these capabilities.[1] With the widespread adoption of electronic health records (EHRs), much of the data necessary to build such tools are already being collected during the course of routine medical care. In order to be clinically useful, such tools should be flexible enough (1) to accommodate the limitations inherent to operational EHR data [Hersh et al., 2014]; (2) to update predictions dynamically as new information becomes available; and (3) to scale to the massive size of modern health records.

We collaborated with Duke Connected Care, the ACO affiliated with the Duke University Health System, to develop predictive tools for chronic kidney disease (CKD). CKD is characterized by a gradual and generally symptomless loss of kidney function over time. CKD and its complications cause poor health, premature death, increased health service utilization, and excess economic costs. CKD is defined and staged by the degree to which a person's estimated glomerular filtration rate (eGFR) is impaired. eGFR is an approximation of overall kidney function and is calculated using a routinely obtained clinical laboratory test (serum creatinine) and demographic information (age, sex and race) [Levey et al., 2009; KDIGO, 2013]. Most clinical laboratories report eGFR automatically with every serum creatinine measurement.

Healthcare providers struggle at many levels to provide optimal care for patients with CKD. First, the majority of healthcare providers fail to recognize the presence of CKD, despite the fact that CKD can be readily identified using simple, eGFR-based laboratory criteria [Szcech et al., 2013; Tuot et al., 2011; Allen et al., 2011]. Second, among those patients with recognized CKD, both primary care providers and kidney specialists struggle to predict which patients will progress to kidney failure (requiring dialysis or kidney transplantation to survive) or suffer from other complications caused by CKD, such as early death from heart attack or stroke [Mendehlsson et al., 2011]. Third, providers often fail to prescribe appropriate preventive treatment to slow disease progression or address com-

[1] http://www.healthcare-informatics.com/article/survey-acos-still-cite-lack-interoperability-biggest-barrier

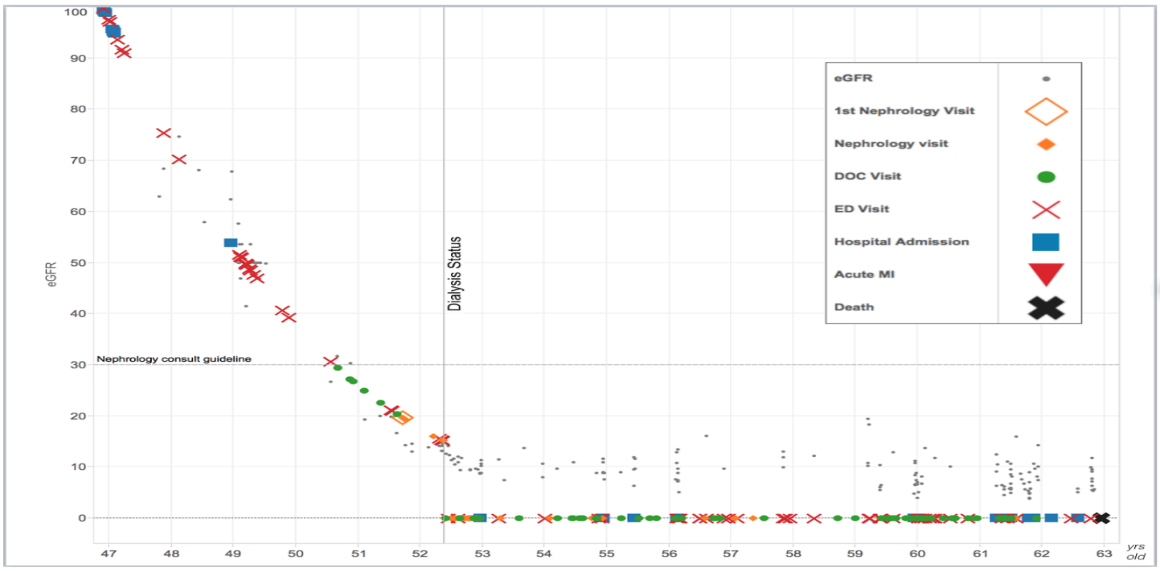

Figure 1: 15-year clinical course of an example patient who experienced both a rapid progression of CKD and a number of other serious health events. Y-axis indicates estimated glomerular filtration rate (eGFR), an estimate of overall kidney function (60-100 is normal, <60 indicates clinically significant kidney disease). X-axis indicates patient age in years. Markers indicate health service use and adverse events. Our model allows us to jointly model progression of CKD, as well as the association between the disease progression and risk for adverse events.

plications [Smart et al., 2014]. Medications such as RAAS drugs can slow progression of CKD if used early enough, while patient counseling and advanced planning can reduce the physical and psychological trauma when kidney failure is imminent.

From a population health management perspective, these characteristics make CKD an ideal condition to model and to develop high-impact care management programs. The challenges surrounding CKD care are best articulated with a representative clinical case, illustrated in Figure 1. A 47 year-old man makes first contact with our health system for emergency treatment of a stroke. His kidney function at this point is normal, although he possesses several risk factors for future CKD. Over the next 5 years, he receives sufficient medical care to detect that his kidney function is deteriorating rapidly (the normal annual rate of kidney function loss at his age is only about 1-2%). His kidney disease goes unnoticed by his healthcare providers, and he does not receive any treatment aimed at slowing progression to total kidney failure. At age 52, he is eventually referred to a kidney specialist, more than a year after his kidney function has fallen below the recommended threshold for such a referral. By this point, kidney failure is inevitable and there is too little time to make advanced preparations for kidney failure, such as pre-emptive kidney transplantation or at-home dialysis. Within 90 days of that first kidney specialist appointment, he develops symptoms of kidney failure and requires hospitalization for emergency dialysis initiation, which is both extremely traumatic and makes him among the most expensive type of patient to treat [Johnson et al., 2015]. He survives on dialysis for about a decade, suffering multiple cardiovascular complications from kidney failure, and ultimately dies at age 63. This patient's story is one of missed opportunities–opportunities that could have been acted upon with accurate predictions using machine learning methods and care management programs.

Our goal is to develop statistical methods that model both the risks of future loss of kidney function and the risks of future complications or adverse health events. The predictions from these models can then be used by healthcare organizations to connect high-risk patients to appropriately targeted interventions. Since the broad aim is to predict which patients will worsen in the near future, we need to model associations between CKD and the multitude of various health outcomes that could occur. CKD frequently coexists with and contributes to cardiovascular disease. In fact, most patients with advanced CKD pass away from cardiovascular complications before the onset of kidney failure. In this article, we choose to focus on two common types of adverse cardiovascular events: heart attacks (acute myocardial infarctions [AMIs]) and strokes (cerebrovascular accidents [CVAs]).

To this end, we develop a joint model that flexibly captures the eGFR trajectory of CKD progression, while simultaneously learning the association between disease trajectory and cardiovascular events. We formulate our approach as a hierarchical latent variable model. Each patient is represented by a set of latent variables characterizing both their disease trajectory and risk of having events. This approach

captures dependencies between the disease trajectory and event risk.

Using our model, we study a large cohort of patients with CKD from the Duke University Health System and make predictions about the trajectory of their disease, as well as their risk of cardiovascular events. Our inference algorithm scales well to the large dataset, and makes accurate predictions that outperform several baselines.

2 PROPOSED JOINT MODEL FOR ELECTRONIC HEALTH RECORDS

In this section, we first describe the structure of electronic health records before introducing our proposed joint model for longitudinal and point process data.

Electronic Health Records

The Duke University Health System's electronic health record (Epic Systems, Madison, WI) stores nearly all available information captured about patients during their encounters within the health system. The EHR contains a large quantity of longitudinal patient data. The vast majority of the data are unstructured, contained within free-text notes and reports. Structured data include demographics, diagnosis and procedural codes, orders, laboratory results, and objective clinical observations (such as vital signs and various nursing assessments). Of particular interest to our work in modeling CKD patients are structured diagnosis codes and laboratory results.

The EHR stores granular information about medical diagnoses using structured, hierarchical codes conforming to ICD-9 (International Classification of Disease, 9th revision), a standardized taxonomy that is used principally for medical billing. For each medical encounter (such as a clinic or emergency department visit), a set of codes is assigned to document the primary problems or diseases that were addressed. In total, there are about 9,000 unique ICD-9 codes. Each clinical diagnosis may have multiple corresponding ICD-9 diagnosis codes. The Agency for Healthcare Research and Quality publishes the Clinical Classifications Software[2], a categorization tool that collapses the thousands of original codes into a few hundred clinically meaningful concepts. We use this mechanism to identify and aggregate codes for CVAs and AMIs, where we use the mean date among all relevant codes within monthly bins to account for multiple codes in a short time period that refer to the same clinical event.

In contrast to diagnosis codes, which capture clinicians' subjective diagnostic impressions, laboratory tests provide objective clinical data. A single medical encounter may include dozens, hundreds or (in the case of hospitalizations) thousands of discrete laboratory test results. Identifying and grouping relevant laboratory test results can be difficult due to lack of standardization and changing conventions over time. For example, serum creatinine, which is a lab test used to calculate eGFR, has more than 18 different names in our EHR that refer to the same value (e.g. "CREA", "Creatinine", "DUAP CREA"). Harmonizing and grouping these lab results required an exhaustive review of laboratory metadata by a subject matter expert.

There are numerous ongoing efforts to develop improved algorithms to identify chronic medical conditions and incident clinical events using a wide assortment of clinical data. Our model is agnostic to the particular algorithm used to identify clinical events. After cleaning and transforming the raw EHR data, we obtained a longitudinal set of eGFRs for each patient, and dates of CVA and AMI diagnoses.

Proposed Model

Our proposed hierarchical latent variable model jointly models longitudinal and point process data by creating different submodels for each type of data, with shared latent variables for each patient inducing dependencies between their two data types. Assume there are N patients, let $\vec{y_i} = \{y_{ij}\}_{j=1}^{N_i}$ denote the N_i observed readings of eGFR for patient i at times $\vec{t_i} = \{t_{ij}\}_{j=1}^{N_i}$, and let $\vec{u_i} = \{u_{ik}\}_{k=1}^{K_i}$ denote the K_i cardiac events patient i experiences (note that K_i may be 0). Let T_i^- be the time patient i is first seen in our sample of their health record, and T_i^+ the final time they are observed. Let z_i, b_i, f_i and v_i be a set of shared hierarchical latent variables for each patient i, to be defined subsequently. Conditioned on these latent variables, to be learned during inference, we make a common conditional independence assumption that the conditional likelihood for patient i factorizes:

$$p(\vec{y_i}, \vec{u_i}|z_i, b_i, f_i, v_i; x_i) = p(\vec{y_i}|z_i, b_i, f_i; x_i) \\ p(\vec{u_i}|z_i, b_i, f_i, v_i; x_i). \tag{1}$$

Longitudinal Submodel

We use a recently proposed model for disease trajectories for our longitudinal submodel [Schulam and Saria, 2015], that was shown to be extremely flexible and accurate at modeling continuous functions of disease progression. Given the set of latent variables for patient i, the longitudinal variables are conditionally independent, i.e. $p(\vec{y_i}|z_i, b_i, f_i) = \prod_{j=1}^{N_i} p(y_{ij}|z_i, b_i, f_i)$. The model assumes each observed longitudinal value is a normally distributed random variable containing a population component, a subpopulation component, an individual component, and a structured noise component:

$$y_i(t) = m_i(t) + \epsilon_i(t),\ \epsilon_i(t) \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \tag{2}$$

$$m_i(t) = \Phi_p(t)^\top \Lambda x_{ip} + \Phi_z(t)^\top \beta_{z_i} + \Phi_l(t)^\top b_i + f_i(t). \tag{3}$$

[2] http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

The first term in (3) is the population component, where $\Phi_p(t) \in \mathbb{R}^{d_p}$ is a fixed basis expansion of time, $\Lambda \in \mathbb{R}^{d_p \times q_p}$ is a coefficient matrix, and $x_{ip} \in \mathbb{R}^{q_p}$ is a vector of baseline covariates.

The second term in (3) is the subpopulation component, where it is assumed person i belongs to latent subpopulation $z_i \in \{1, \ldots, G\}$. Each subpopulation is associated with a unique disease trajectory represented using B-splines, in particular, $\Phi_z(t) \in \mathbb{R}^{d_z}$ is a fixed B-spline basis expansion of time with $\beta_g \in \mathbb{R}^{d_z}$ the coefficient vector for group g. We assign z_i a multinomial logistic regression prior that depends on baseline covariates $x_{iz} \in \mathbb{R}^{q_z}$: $p(z_i = g) \propto \exp\{w_g^\top x_{iz}\}$, where $\{w_g\}_{g=1}^G$ are regression coefficients with $w_1 \equiv 0$ for identifiability.

The third term is the individual component, allowing for individual-specific long-term deviations in trajectory that are learned dynamically as more data is available. $\Phi_l(t) \in \mathbb{R}^{d_l}$ is a fixed basis expansion of time, and $b_i \in \mathbb{R}^{d_l}$ is a random effect for patient i, with prior $b_i \sim N(0, \Sigma_b)$.

Finally, $f_i(t)$ is the structured noise process that captures transient trends in disease trajectory. This is modeled using a zero-mean Gaussian process with Ornstein-Uhlenbeck covariance function $K_{OU}(t_1, t_2) = \sigma_f^2 \exp\{-\frac{|t_1 - t_2|}{l}\}$. This kernel is well-suited for this task, as it is mean-reverting and has no long-range dependence between deviations [Schulam and Saria, 2015].

Point Process Submodel

We choose to model the times $\vec{u_i} = \{u_{ik}\}_{k=1}^{K_i}$ that a person has an adverse event as a Poisson process. A common choice for the rate function from related literature in survival analysis corresponds to the hazard function from the Cox proportional hazards model. We make this choice in this work, for reasons both of simplicity and also computational efficiency as we discuss later. The conditional likelihood for the Poisson process for patient i on the interval $[T_i^-, T_i^+]$, with events at times $\{u_{ik}\}_{k=1}^{K_i}$, is given by:

$$p(\vec{u_i} | z_i, b_i, f_i, v_i) = \prod_{k=1}^{K_i} r_i(u_{ik}) \exp\{-\int_{T_i^-}^{T_i^+} r_i(t) dt\}, \tag{4}$$

where we specify the rate function for patient i as:

$$r_i(t) = r_0(t) \exp\{\gamma^\top x_{ir} + \alpha m_i(t) + \delta m_i'(t) + v_i\}. \tag{5}$$

We assume that $r_0(t)$ is a piecewise constant function with jumps at fixed quantiles of the event times, and heights $\{a_l\}_{l=1}^{N_r}$. The parameter $\gamma \in \mathbb{R}^{q_r}$ specifies the association between baseline covariates $x_{ir} \in \mathbb{R}^{q_r}$ and the risk for an event, while parameters α and δ specify the association between the risk for an event and the expected mean and expected slope of the longitudinal variable at that time, respectively.[3] Finally, the latent variable v_i, with prior $v_i \sim N(0, \sigma_v^2)$, represents an additional random effect (called a frailty term in survival analysis), multiplicatively adjusting an individual's overall risk for events. In order to compute the likelihood, we must compute the definite integral in (4) numerically. We find that the trapezoid rule works fine, although other options such as Gaussian quadrature are also possible.

3 RELATED WORK

There is a rich literature, mostly from biostatistics, on joint models typically for longitudinal data and time-to-event data with right censoring. See [Rizopoulos, 2012] for a thorough introduction to these types of joint models. A slightly different flavor of joint models is presented in [Proust-Lima et al., 2014]. These models differ in that instead of the longitudinal value directly influencing the event rate, they consider latent subpopulations of individuals within which it is assumed there is a different average profile of both the longitudinal value and risk of the event.

Most directly relevant to our work are several methods for modeling longitudinal data and recurrent event data [Liu and Huang, 2009; Kim et al., 2012; Han et al., 2007]. However, these methods share several notable weaknesses. First, the form for their longitudinal models are simplistic, all being mixed effects models. Such models are inflexible and will fail to capture the types of trajectories that our model can, through its mixture model and both long and short-term individual-specific deviations. In addition, these works as well as most of the literature on joint models rely on computationally expensive inference algorithms, thereby limiting their use to small datasets. Typically EM or gradient methods are employed for Maximum Likelihood Estimation, or MCMC in Bayesian settings. It is extremely uncommon to find a published joint model applied to a dataset of over 1000 individuals. However, our scalable variational inference algorithm, developed in the next section, is much more efficient, facilitating use in large-scale applications where there can be tens or even hundreds of thousands of patients.

Within the medical literature, there have been numerous studies on predicting adverse events such as kidney failure, death, or cardiac events in patients with CKD; for instance, [Tangri et al., 2011] is a common reference. In almost every case, the models developed are Cox proportional hazards models for time-to-event data, or logistic regression models for occurrence of an event in a specified time window. As such, these models are all static and use only a single snapshot of patient data to make predictions, which precludes the ability to generate dynamic predictions.

[3] Since $f_i(t)$ with an OU kernel is not differentiable, we let $m_i'(t)$ be the sum of the slopes of the first three terms in (3).

In recent years there has been much interest in machine learning in modeling electronic health records and other forms of healthcare data. For instance, [Lian et al., 2015] use hierarchical point processes to predict hospital admissions, and [Ranganath et al., 2015] develop a dynamic factor model to learn relationships between diseases and predict future diagnosis codes. Closest to our work in the application is [Perotte et al., 2015], who explore using time-series models to predict a time-to-event (progression from CKD stage 3 to stage 4) in CKD patients.

4 INFERENCE

As with most complex probabilistic generative models, the computational problem associated with fitting the model is estimation of the posterior distribution of latent variables and model parameters given the observed data. Exact computation of the posterior is intractable, and requires approximation to compute. To this end, we develop a mean field variational inference [Jordan et al., 1999] algorithm to approximate the posterior distribution of interest.

Variational methods transform the task of posterior inference into an optimization problem. The optimization problem posed by variational inference is to find a distribution q in some approximating family of distributions that is close in KL divergence to the true posterior. Equivalently, the problem can be viewed as maximizing what is known as the evidence lower bound (ELBO) [Bishop, 2006]:

$$\mathcal{L}(q) = E_q[\log p(y, u, z, b, f, v, \Theta) - \log q(z, b, f, v, \Theta)], \tag{6}$$

which forms a lower bound on the marginal likelihood $p(y, u)$ of our model.

Variational Approximation

Recall for our model that the model parameters are $\Theta = \{\Lambda, W, \beta, a, \gamma, \alpha, \delta\}$, and the local latent variables specific to each person are their subpopulation assignment z_i, random effects b_i and v_i, and structured noise function f_i. The joint distribution for our model can be expressed as:

$$\begin{aligned} p(y, u, z, b, f, v, \Theta) = p(\Theta) \prod_{i=1}^{N} \prod_{j=i}^{N_i} p(y_{ij} | z_i, b_i, f_i(t_{ij}), \Theta) \\ p(\vec{u_i} | z_i, b_i, f_i, v_i, \Theta) p(z_i) p(b_i) p(f_i) p(v_i) \end{aligned} \tag{7}$$

We make the mean field assumption for the variational distribution, which assumes that in the approximate posterior q, all the latent variables are independent. This implies that $q(z, b, f, v, \Theta) = q(\Theta) \prod_{i=1}^{N} q_i(z_i, b_i, f_i, v_i)$, where:

$$\begin{aligned} q_i(z_i, b_i, f_i, v_i) = q_i(z_i | \nu_{z_i}) q_i(b_i | \mu_{b_i}, \Sigma_{b_i}) \\ q_i(v_i | \mu_{v_i}, \sigma^2_{v_i}) q_i(f_i). \end{aligned} \tag{8}$$

The assumed variational distributions for z_i, b_i, and v_i are the same family as their prior distribution, i.e. multinomial, multivariate normal, and univariate normal. For the variational form for f_i, we adapt ideas from the variational learning for sparse GPs literature [Lloyd et al., 2014; Titsias, 2009] to approximate the true posterior over f_i. In order to evaluate the ELBO in (6), we will need to evaluate $E_{q_i}[f_i]$ at times $\vec{t_i}$ for the longitudinal likelihood, as well as at $\vec{u_i}$ and at a grid of times t_i^{grid} for the point process likelihood (the grid is for the numerical integration). We choose to treat the observed observation times $\vec{t_i}$ as pseudo-inputs; this helps reduce overfitting and reduces the number of variational parameters to learn. In particular:

$$\begin{aligned} q_i(f_i(\vec{t_i}), f_i(\vec{u_i}), f_i(t_i^{\text{grid}})) = p(f_i(\vec{u_i}), f_i(t_i^{\text{grid}}) | f_i(\vec{t_i})) \\ q(f_i(\vec{t_i}) | \mu_{f_i}, \Sigma_{f_i}). \end{aligned} \tag{9}$$

We allow a free-form multivariate Gaussian distribution for f_i at the longitudinal observation times, and use a so-called conditional Gaussian process for the distribution at $\vec{u_i}, t_i^{\text{grid}}$, i.e. the true conditional distribution of the joint multivariate normal, $f_i | f_i(\vec{t_i}) \sim \mathcal{GP}(\mu(t), \Sigma(t, t'))$:

$$\mu(t) = K_{t,\vec{t_i}} K^{-1}_{\vec{t_i},\vec{t_i}} f_i(\vec{t_i}) \tag{10}$$

$$\Sigma(t, t') = K_{t,t'} - K_{t,\vec{t_i}} K^{-1}_{\vec{t_i},\vec{t_i}} K_{\vec{t_i},t'} \tag{11}$$

where $K_{t,\vec{t_i}}, K_{\vec{t_i},\vec{t_i}}, K_{t,t'}$ are matrices evaluated at t, t', and $\vec{t_i}$ using the OU covariance kernel from Section 2.

Although priors on the model parameters Θ may be imposed, i.e. log-normal on a and normal on the rest, in our work we learn their maximum likelihood estimate (MLE) instead, and let $q(\Theta)$ be a delta function. Thus, the goal of our variational algorithm is to learn optimal variational parameters $\lambda_i = \{\nu_{z_i}, \mu_{b_i}, \Sigma_{b_i}, \mu_{v_i}, \sigma^2_{v_i}, \mu_{f_i}, \Sigma_{f_i}\}$ for each individual i, as well as a point estimate $\hat{\Theta}$ for the model parameters. In practice, we optimize the Cholesky decompositions L_{b_i}, L_{f_i} for the covariance matrices $\Sigma_{b_i}, \Sigma_{f_i}$.

Solving the Optimization Problem

In traditional settings for variational inference, the objective function is iteratively optimized by maximizing the variational parameters associated with each latent variable or parameter, holding the rest fixed. In models where the log complete conditional distributions (log of the conditional distribution of each latent variable given everything else) have analytic expectations with respect to the variational approximation, closed form EM-style updates are available for the variational parameters. This convenient property is typically observed in conditionally conjugate models, where each log complete conditional will be in the exponential family [Ghahramani and Beal, 2001].

Recently there has been much interest in applying variational methods to more complex models that do not ex-

hibit conjugacy. In many cases, it is intractable to even evaluate the ELBO analytically, since one or both of the expectations in (6) have no closed form. In these cases, variational algorithms have been developed that rely on sampling from the variational approximation [Ranganath et al.,2014; Rezende et al., 2014]. However, because of the form we chose for r_i, it is possible to calculate a closed form approximation to the ELBO for our model (approximate due to the numerical integration; see Appendix for details). As such, we use the automatic differentiation package *autograd* [4] in Python to compute analytic gradients in order to optimize the bound. At each iteration of the algorithm, we optimize the local variational parameters in parallel using exact gradients. To optimize the global parameters, we turn to stochastic optimization.

Stochastic optimization has become a commonly used tool in variational inference. Rather than using every single observation to compute the gradient of the ELBO with respect to Θ, we can compute a noisy gradient based on a sampled batch of observations [Hoffman et al., 2013]. As long as the noisy gradient is unbiased and the learning rate ρ_t at each iteration satisfies the Robbins Monro conditions ($\sum_{t=1}^{\infty} \rho_t = \infty,\ \sum_{t=1}^{\infty} \rho_t^2 < \infty$), the stochastic optimization procedure will converge to a local maximum. To set the learning rate we use the AdaGrad algorithm, which adaptively allows for a different learning rate for each parameter. The learning rate for each parameter is scaled by the square root of a running sum of the squares of historical gradients [Duchi et al., 2011].

Algorithm

Algorithm 1 summarizes the procedure to learn an approximate posterior for the local latent variables and a point estimate for the model parameters.

Data: data y, u; hyperparameters.
Result: point estimate $\hat{\Theta}$, approximate posteriors q_i.
Initialize global parameters Θ.
repeat
  Randomly sample data for S patients, $\{y_s, u_s\}_{s=1}^{S}$.
  for *s = 1:S in parallel* **do**
    Optimize local variational parameters for q_s via gradient ascent.
  end
  Compute the noisy gradient for Θ.
  Update Θ using AdaGrad.
until *convergence of the ELBO*;
Algorithm 1: Stochastic Variational Inference algorithm for our Joint Model.

[4] https://github.com/HIPS/autograd

5 EMPIRICAL STUDY

In this section we describe our experimental setup and results on our real dataset.

Dataset

Our dataset comprises longitudinal and cardiac event data from 23,450 patients with stage 3 CKD or higher within our university health system. IRB approval (#Pro00066690) was obtained for this work. We first created an initial cohort of roughly 600,000 patients that had at least one encounter in the health system in the year prior to Feb. 1, 2015. This includes all types of encounters within the health system, including inpatient, outpatient, and emergency department visits. From this, we filtered to patients who had at least ten recorded values for serum creatinine, the laboratory value required to calculate eGFR. We next filtered to patients that had Stage 3 CKD or higher, indicative of moderate to severe kidney damage, defined as two eGFR measurements less than 60 mL/min separated by at least 90 days. Finally, since the recorded eGFR values are extremely noisy and eGFR is only a valid estimate of kidney function at steady state, we take the mean of eGFR readings in monthly time bins for each patient. Rapid fluctuations in acute illness are related to long term risk, but we have not yet explicitly incorporated this into our modeling.

After this preprocessing, on average each patient has 22.9 eGFR readings (std dev 13.6; median 19.0). In order to align the patients on a common time axis, for each patient we fix $t = 0$ to be their first recorded eGFR reading below 60 mL/min. The adverse events of interest in our experiments are AMIs and CVAs, and these were identified using ICD-9 codes as detailed in Section 2. 13.4% of patients had at least one code for AMI (among those with at least one: mean 4.1, std dev 7.1, median 2.0), and likewise 17.4% of patients had at least one code for CVA (mean 6.4, std dev 13.3, median 3.0). We use the same set of baseline covariates for x_{ip}, x_{iz}, x_{ir}: baseline age, race and gender, and indicator variables for hypertension and diabetes. Note that x_{ip}, x_{iz} include an intercept while x_{ir} does not.

For the experiments, we used ten fold cross validation with training sets of 21,105 patient records and test sets of 2,345 records. We fit separate joint models for CVA events and AMI events.

Evaluation Metrics

After learning a point estimate for the global model parameters during training, they are held fixed. Then, an approximate posterior is fit to each patient in the test set, where we allow the learning algorithm to see the first 60% of a patient's eGFR trajectory (and any events before then) and hold out the remaining 40% (and future events). Predictions about future disease trajectory and adverse events are

made by drawing samples from the approximate posterior predictive distribution.

We evaluate our model on two tasks to asses predictive performance of each submodel. For the longitudinal submodel, we compute the mean squared error (MSE) and mean absolute error (MAE) for predictions about held-out eGFR values. For the point process submodel, we view the problem of predicting whether any event will occur in a given future time window (in our experiments, 1-5 years) as a binary classification problem. We report the area under the ROC curve (AUROC) and area under the precision-recall curve (AUPR) as evaluation metrics for each binary classification task. Calculating the probability of an event in a future time window $[T_i, T_i + c]$ for person i is easily computed as $1 - \exp\{-\int_{T_i}^{T_i+c} r_i(t)dt\}$.

Baselines

For the longitudinal submodel, we compare against the model in [Schulam and Saria, 2015], since we use their model as our longitudinal submodel. However, because our model was trained jointly with the point process submodel we do not in general learn the same model parameters, since the parameters for the learned trajectories are also influenced by the event data.

For the point process submodel, we compare against two standard baselines. The first is a simple Cox proportional hazards model from survival analysis, where we use the same set of time independent covariates x_{ir} as in our model. The likelihood is the same as (4), but now $r_i(t) = r_0(t) \exp\{\gamma^\top x_{ir}\}$. We also compare against a Cox model with time-dependent covariates, where $r_i(t) = r_0(t) \exp\{\gamma^\top x_{ir} + \alpha y_i(t)\}$, with $y_i(t)$ a step function denoting the most recent observed eGFR up until time t. Due to the lack of scalable inference algorithms for related works from the joint modeling literature, we were unable to compare against them on our large patient cohort.

Hyperparameters

We learn point estimates for hyperparameters $\sigma_\epsilon, \Sigma_b, \sigma_v, \sigma_f, l$ by maximizing the ELBO with respect to them. Additional hyperparameters include G, N_r, and the choice of basis expansions Φ_p, Φ_z, Φ_l in the longitudinal submodels. We let Φ_p and Φ_l be linear basis functions of time, thus allowing for population covariates and individual heterogeneity to shift the intercept and slope of eGFR trajectory. We let Φ_z be a B-spline expansion of time with degree two and twelve knots at equally spaced quantiles of eGFR observation times. We fix $G = 15$ and $N_r = 9$. Finally, we set the global scale parameter for AdaGrad to 0.1, and subsample 250 observations at a time. We experimented with other values for these fixed hyperparameters without major changes in performance.

Results

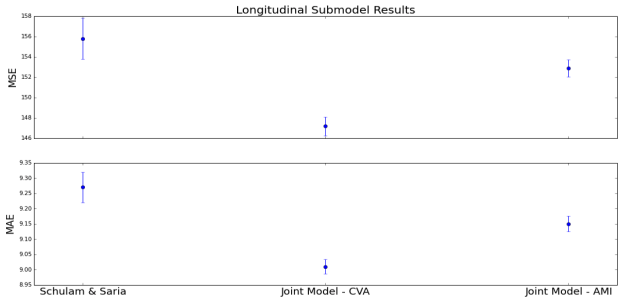

Figure 2: Mean MSE and MAE from longitudinal submodels. Error bars are one standard error.

Figure 2 highlights the results from the longitudinal submodel, where we present the mean MSEs and MAEs across the test sets. The longitudinal submodel from our joint model performs slightly better than the method of [Schulam and Saria, 2015] fit independently to the eGFR values. Figure 3 highlight the results from the point process submodel. Our proposed joint model performs substantially better than the two baselines at predicting future events, in terms of both AUROC and AUPR.

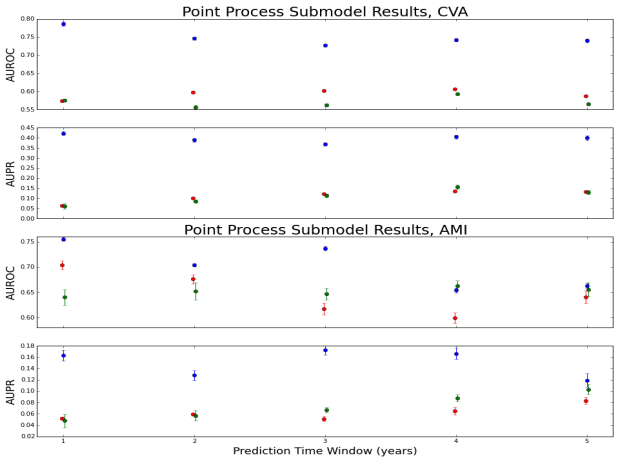

Figure 3: Mean AUROC and AUPR for CVA and AMI events. Blue is proposed Joint Model, red is Cox, green is time-varying Cox. Error bars are one standard error.

In addition, in this dataset it appears that prediction of CVA events is slightly easier than prediction of AMIs. For the CVA joint model, we estimate that $\alpha = -0.063$ and $\delta = -0.061$ while for the AMI joint model, $\alpha = -0.158$ and $\delta = -0.069$ (standard errors for all four estimates < 0.01, from the cross validation). The signs of these parameters agree with clinical intuition that patients with lower overall eGFR values and more rapid eGFR declines should be at higher risk for adverse events. It appears there is a slightly stronger association between eGFR trajectory and risk for AMIs compared to CVAs.

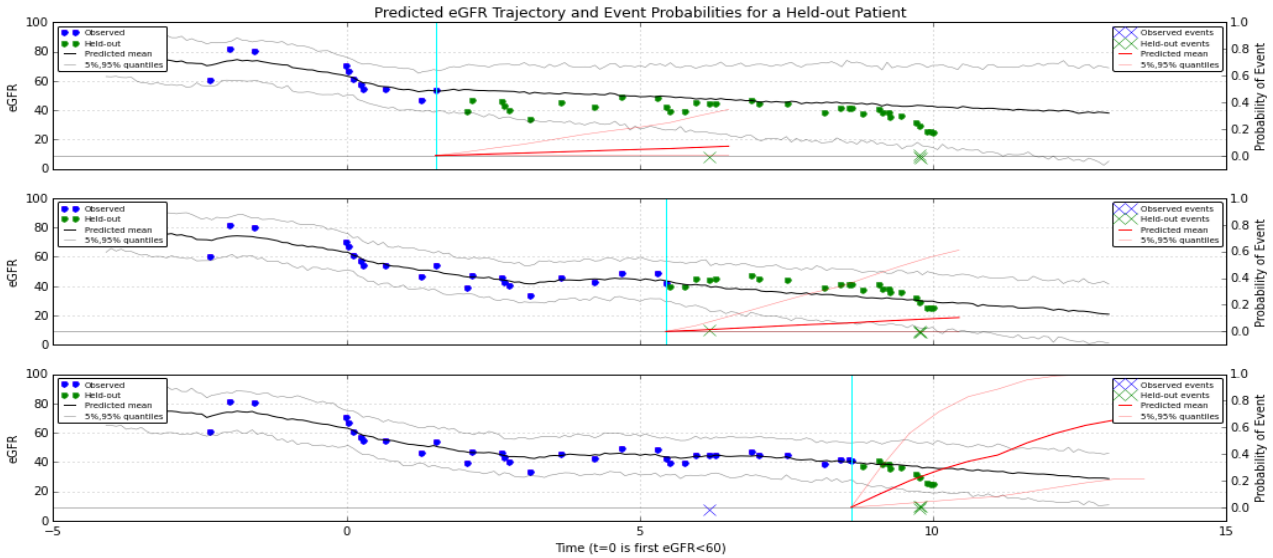

Figure 4: Dynamic predictions from our joint model. In each row, the parameters for this individual are refit as more data is made available (information to the left of the light blue lines is used to refit parameters). Blue circles and x's correspond to observed eGFR readings and CVA events, while green correspond to yet-unseen data.

Figure 4 shows an example of dynamic predictions over time for a test patient. In the three rows of the figure, we make predictions about the test patient after observing the first 25%, 50% and 75% of their disease trajectory and adverse events (in this example, CVAs). For each row we relearn the patient's parameters using information to the left of the vertical light blue line. As we observe more data, the longitudinal model updates its prediction about future disease trajectory and provides a reasonable forecast for the steady decline of this patient's eGFR. In the second row, as the model sees that the patient's trajectory is decreasing faster than in the first row, it correspondingly increases the probability of a future event. In the third row, after the model sees the patient's first CVA event, it further increases the probability of a future event.

6 DISCUSSION

In this paper, we have proposed a new joint model for longitudinal and point process data, and applied it to disease trajectory modeling and prediction of adverse events in patients with chronic kidney disease. We developed the first variational inference algorithm for this class of models, allowing us to fit our model to a large set of longitudinal patient data that is over an order of magnitude the size of datasets used by related methods. We find that our model yields good performance on the tasks of predicting future kidney function and predicting cardiovascular events.

Although our work is a promising first step for developing predictive models from EHR data and applying them to real clinical tasks, there are numerous inherent limitations to EHR data [Hersh et al., 2014]. Data quality is often poor, complicated by inaccurate, inconsistent or missing information. The EHR at a single organization may fail to capture the full patient story and all relevant outcomes of interest, as is the case when patients receive care from multiple, non-interoperable healthcare systems over time. Relevant patient reported outcomes, such as perceived quality of life, are rarely captured by EHRs. Events such as death may not be registered, particularly when patients die outside of the hospital. Data may be biased; certain laboratory tests may be performed only when a clinician suspects an abnormality. Furthermore, many clinical data are collected for billing purposes rather than patient care or research, distorting the relative importance of certain elements.

There are many directions in which we plan to extend this work. Future models will be multivariate in both longitudinal markers and in event processes. Inclusion of additional longitudinal variables such as blood pressure, albuminuria, and hemoglobin A1c will be important, since these are well known to be clinically important for monitoring cardiovascular and kidney health. Jointly modeling multiple event processes will allow us to learn correlations between different types of events. More flexible models, particularly for the event processes, should improve model performance, for instance using Gaussian Process modulated Poisson processes or Hawkes processes instead of employing the proportional hazards assumption as we do in this work. By further refining and deploying a flexible, scalable model such as ours, ACOs around the country can intervene on high-risk patients and realize the potential benefits of precision medicine.

Acknowledgements

Joseph Futoma is supported by an NDSEG fellowship. Dr. Cameron is supported by a Duke Training Grant in Nephrology (5T32DK007731). This project was funded by both the Duke Translational Research Institute and the Duke Institute for Health Innovation. Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the NIH under Award Number UL1TR001117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

7 APPENDIX

We present the full derivation of the ELBO for our model. We can rewrite the expression for the ELBO in (6) as:

$$\mathcal{L}(q) = \sum_{i=1}^{N} \mathcal{L}(q_i), \tag{12}$$

$$\begin{aligned}\mathcal{L}(q_i) = &E_{q_i}[\log p(\vec{y_i}|z_i, b_i, f_i, \Theta) + \log p(\vec{u_i}|z_i, b_i, f_i, v_i, \Theta)] \\ &- KL(q_i(b_i)||p(b_i)) - KL(q_i(v_i)||p(v_i)) \\ &- KL(q_i(z_i)||p(z_i)) - KL(q_i(f_i(\vec{t_i}))||p(f_i(\vec{t_i}))).\end{aligned} \tag{13}$$

Computation of the KL divergence terms are standard. We focus our attention on the first two terms in (13).

The first term in (13) is the variational expectation of the log likelihood for the longitudinal submodel. To compute this, we need to calculate $E_{q_i}[(\vec{y_i} - m_i(\vec{t_i}))^\top(\vec{y_i} - m_i(\vec{t_i}))]$. It is straightforward to expand this product and calculate the expectation of each term. The relevant expectations are:

$$E_{q_i(z_i)}[\beta_{z_i}] = \sum_{g=1}^{G} \nu_{z_i,g}\beta_g \tag{14}$$

$$E_{q_i(b_i)}[b_i] = \mu_{b_i} \tag{15}$$

$$E_{q_i(f_i)}[f_i(\vec{t_i})] = \mu_{f_i} \tag{16}$$

$$\begin{aligned}&E_{q_i(z_i)}[(\Phi_z(\vec{t_i})\beta_{z_i})^\top(\Phi_z(\vec{t_i})\beta_{z_i})] = \\ &\sum_{g=1}^{G} \nu_{z_i,g}(\Phi_z(\vec{t_i})\beta_g)^\top(\Phi_z(\vec{t_i})\beta_g)\end{aligned} \tag{17}$$

$$\begin{aligned}&E_{q_i(b_i)}[(\Phi_l(\vec{t_i})b_i)^\top(\Phi_l(\vec{t_i})b_i)] = \\ &\text{Tr}(\Phi_l(\vec{t_i})\Sigma_{b_i}\Phi_l(\vec{t_i})^\top) + \mu_{b_i}^\top\Phi_l(\vec{t_i})^\top\Phi_l(\vec{t_i})\mu_{b_i}\end{aligned} \tag{18}$$

$$E_{q_i(f_i(\vec{t_i}))}[f_i(\vec{t_i})^\top f_i(\vec{t_i})] = \text{Tr}(\Sigma_{f_i}) + \mu_{f_i}^\top\mu_{f_i} \tag{19}$$

The second term in (13) is the variational expectation of the log likelihood for the point process submodel:

$$\begin{aligned}&E_{q_i}[\log p(\vec{u_i}|z_i, b_i, f_i, v_i, \Theta)] = \\ &E_{q_i}[\sum_{k=1}^{K} \log r_i(u_{ik}) - \int_{T_i^-}^{T_i^+} r_i(t)dt].\end{aligned} \tag{20}$$

Each term in the summation in (20) is given by:

$$\begin{aligned}E_{q_i}[\log r_i(u_{ik})] = &\log r_0(u_{ik}) + \gamma^\top x_{ir} + \alpha E_{q_i}[m_i(u_{ik})] \\ &+ \delta E_{q_i}[m_i'(u_{ik})] + E_{q_i}[v_i],\end{aligned} \tag{21}$$

where $E_{q_i}[v_i] = \mu_{v_i}$ and $E_{q_i}[m_i(u_{ik})], E_{q_i}[m_i'(u_{ik})]$ are simple to compute using (14) and (15), where we use the time derivatives of the bases $\Phi_p', \Phi_z', \Phi_l'$ in place of the actual bases for the latter. The only nontrivial term in $E_{q_i}[m_i(u_{ik})]$ is $E_{q_i}[f_i(u_{ik})]$. However, due to conjugacy, we have that for arbitrary t:

$$\begin{aligned}q_i(f_i(t)) &= \int p(f_i(t)|f_i(\vec{t_i}))q_i(f_i(\vec{t_i})) \\ &\equiv \mathcal{GP}(f_i; \mu(t), \Sigma(t, t'))\end{aligned} \tag{22}$$

$$\mu(t) = K_{t,\vec{t_i}}K_{\vec{t_i},\vec{t_i}}^{-1}\mu_{f_i} \tag{23}$$

$$\begin{aligned}\Sigma(t,t') = &K_{t,t'} - K_{t,\vec{t_i}}K_{\vec{t_i},\vec{t_i}}^{-1}K_{\vec{t_i},t'} \\ &+ K_{t,\vec{t_i}}K_{\vec{t_i},\vec{t_i}}^{-1}\Sigma_{f_i}K_{\vec{t_i},\vec{t_i}}^{-1}K_{\vec{t_i},t'},\end{aligned} \tag{24}$$

so we can use (23) to compute $E_{q_i}[f_i(u_{ik})]$; the K matrices are the OU kernel evaluated at the relevant times.

The final term to compute is the integral in (20). Since we approximate it numerically, we need to evaluate $E_{q_i}[r_i(t)]$ for arbitrary times t:

$$E_{q_i}[r_i(t)] = r_0(t)e^{\gamma^\top x_{ir}}E_{q_i}[e^{\alpha m_i(t) + \delta m_i'(t) + v_i}]. \tag{25}$$

Using the mean field assumption, this expectation of products factorizes into products of expectations. There are no local latent variables corresponding to the population term in $m_i(t)$, so that term can be brought outside the expectation. Since $q_i(v_i) \sim N(\mu_{v_i}, \sigma_{v_i}^2)$ we have that e^{v_i} is log-normal, hence $E_{q_i}[e^{v_i}] = e^{\mu_{v_i} + \frac{\sigma_{v_i}^2}{2}}$. Expanding $m_i(t)$ and $m_i'(t)$ in (25) leads to three final expectations to compute. The first is:

$$E_{q_i(z_i)}[e^{(\alpha\Phi_z(t) + \delta\Phi_z'(t))^\top\beta_{z_i}}] = \sum_{g=1}^{G} \nu_{z_i,g}e^{(\alpha\Phi_z(t) + \delta\Phi_z'(t))^\top\beta_g} \tag{26}$$

The last two are $E_{q_i}[e^{(\alpha\Phi_l(t)^\top + \delta\Phi_l'(t)^\top)b_i}]$ and $E_{q_i}[e^{\alpha f_i(t)}]$. Since the variational distributions for b_i and $f_i(t)$ are multivariate normal (from (8) and (22)-(24)), the exponential of an affine transformation of them will be multivariate log-normal. We can use this with the fact that if $X \sim N(\mu, \Sigma)$ is multivariate normal, then $Y = e^X$ is multivariate log-normal with mean $E[Y]_i = e^{\mu_i + \frac{\Sigma_{ii}}{2}}$ to compute the desired variational expectations.

To compute noisy gradients of the ELBO with respect to Θ, we randomly sample S observations $\{y_s, u_s\}_{s=1}^S$ at each iteration, and compute the gradient of $\hat{\mathcal{L}}(q) \equiv \frac{N}{S}\sum_{s=1}^S \mathcal{L}(q_s)$, which equals $\mathcal{L}(q)$ in expectation.

References

A. Allen, J. Forman, et al. Primary Care Management of Chronic Kidney Disease. *Journal of General Internal Medicine*, 26(4):386-392, 2011.

D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, 33(7):1123-1131, 2014.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York., 2006.

J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12:2121-2158, 2011.

Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. *NIPS*, 2001.

J. Han, E. H. Slate, and E. A. Pena. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Stat Med* 26(29): 5285-5302, 2007.

W. Hersh, M. Weiner et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care* 51: S30-S37, 2013.

M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14:1303-1347, 2013.

T. Johnson, D. Rinehart, J. Durfee, et al. For Many Patients Who Use Large Amounts of Health Care Services, the Need Is Intense Yet Temporary. *Health Affairs*, 34(8): 1312-1319, 2015.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183-233, 1999.

KDIGO 2012. Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Journal of the International Society of Nephrology*, (3)1, 2013.

S. Kim, D. Zeng, L. Chambless, and Y. Li. Joint Models of Longitudinal and Recurrent Events with Informative Terminal Event. *Stat Biosci* 4(2): 262-281, 2012.

A.S. Levey, L. A. Stevens, et al. A New Equation to Estimate Glomerular Filtration Rate. *Annals of internal medicine*,150(9):604-612, 2009.

W. Lian, R. Henao, V. Rao, J. Lucas, and L. Carin. A Multitask Point Process Predictive Model. *ICML*, 2015.

L. Liu and X. Huang. Joint Analysis of Correlated Repeated Measures and Recurrent Events Processes in the Presence of Death, With Application to a Study on Acquired Immune Deficiency Syndrome. *JRSS, C* 58(1):65-81, 2009.

C. Lloyd, T. Gunter, M. A. Osborne, and S. J. Roberts. Variational inference for Gaussian process modulated Poisson proceses. *ICML*, 2015.

D. Mendelssohn, B. Curtis, K. Yeates, et al. Suboptimal initiation of dialysis with and without early referral to a nephrologist. *Nephrology Dialysis Transplantation*, 26(9):2859-65, 2011.

R. B. Parikh, M. Kakad, and D. W. Bates. Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. *JAMA*, 315(7):651-652, 2016.

A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 2015.

C. Proust-Lima, M. Sene, J. Taylor, H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* 23(1):74-90, 2014.

R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. The Survival Filter: Joint Survival Analysis with a Latent Time Series. *UAI*, 2015.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. *AISTATS*, 2014.

D. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall / CRC Biostatistics Series, 2012.

P. Schulam and S. Saria. A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure. *NIPS*, 2015.

N. Smart, G. Dieberg, M. Ladhanni, and T. Titus. Early referral to specialist nephrology services for preventing the progression to end-stage kidney disease. *Cochrane Database of Systematic Reviews*, 18(6), 2014.

L. Szczech, R. Stewart, H. Su, et al. Primary Care Detection of Chronic Kidney Disease in Adults with Type-2 Diabetes: The ADD-CKD Study. *PLoS ONE*, 9(11), 2014.

N. Tangri, L. A. Stevens, et al. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA* 305(15):1553-1559, 2011.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. *AISTATS*, 2009.

D. Tuot, L. Plantinga, C. Hsu, et al. Chronic Kidney Disease Awareness Among Individuals with Clinical Markers of Kidney Dysfunction. *Clinical Journal of the American Society of Nephrology*, 6(8):1838-1844, 2011.